Adversarial Patch Attack for Ship Detection via Localized Augmentation

Chun Liu, Panpan Ding, Zheng Zheng, Hailong Wang, Bingqian Zhu, Tao Xu, Zhigang Han, Jiayao Wang

Abstract-Current ship detection techniques based on remote sensing imagery primarily rely on the object detection capabilities of deep neural networks (DNNs). However, DNNs are vulnerable to adversarial patch attacks, which can lead to misclassification by the detection model or complete evasion of the targets. Numerous studies have demonstrated that data transformationbased methods can improve the transferability of adversarial examples. However, excessive augmentation of image backgrounds or irrelevant regions may introduce unnecessary interference, resulting in false detections of the object detection model. These errors are not caused by the adversarial patches themselves but rather by the over-augmentation of background and non-target areas. This paper proposes a localized augmentation method that applies augmentation only to the target regions, avoiding any influence on non-target areas. By reducing background interference, this approach enables the loss function to focus more directly on the impact of the adversarial patch on the detection model, thereby improving the attack success rate. Experiments conducted on the HRSC2016 dataset demonstrate that the proposed method effectively increases the success rate of adversarial patch attacks and enhances their transferability.

Index Terms—Adversarial attack, Adversarial patch, Local data augmentation, Ship imagery, Object detection, Black-box attack.

I. INTRODUCTION

Ith the rapid development of remote sensing technology, the volume of imagery and video collected by Earth observation satellites and aerial platforms has grown significantly. Detecting ship targets from these images is of great significance for sea traffic monitoring, ship search and rescue, fisheries management, and maritime situational awareness [1].

Deep neural networks (DNNs) now have shown impressive performance for ship detection [2], providing effective tools

Chun Liu is with the State Key Laboratory of Spatial Datum, College of Remote Sensing and Geoinformatics Engineering, Faculty of Geographical Science and Engineering, Henan University, Zhengzhou 450046, China, and the National Key Laboratory of Integrated Technology on Aircraft Control, Beihang University, Beijing 100191, China. Panpan Ding, Hailong Wang, and Bingqian Zhu are with the School of Computer and Information Engineering, Henan University, Kaifeng 475004, China. Zheng Zheng is with the National Key Laboratory of Integrated Technology on Aircraft Control, School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. Tao Xu is with the State Key Laboratory of Spatial Datum, Henan University, Zhengzhou 450046, China, and the School of Computer and Information Engineering, Henan University, Kaifeng 475004, China. Zhigang Han and Jiayao Wang are with the State Key Laboratory of Spatial Datum, College of Remote Sensing and Geoinformatics Engineering, Faculty of Geographical Science and Engineering, Henan University, Zhengzhou 450046, China. (email:liuchun@henu.edu.cn;1823373789@qq.com;zhengz@buaa.edu.cn; whl0032@gq.com;104754231411@henu.edu.cn;txu@henu.edu.cn; zghan@henu.edu.cn;)

for feature extraction from remote sensing images. However, DNNs are vulnerable to adversarial attacks which have raised widespread concern. Adversarial attacks aim to mislead models by introducing small perturbations to inputs [3]. Based on the attacker's knowledge, adversarial attacks can be classified as white-box (full access to the model) or black-box (interaction through input-output only). Since most deployed systems conceal internal structures, black-box attacks present more realistic threats. They often rely on the transferability of adversarial examples—input samples crafted on a surrogate model that can deceive other models.

Adversarial attacks can be also categorized into pixel-based and patch-based types. Pixel-based attacks modify part or all of an image's pixels [4], often remaining imperceptible to humans [5]. Xie et al. [6] proposed DAG, the first adversarial method for object detection. Li et al. [7] introduced RAP, which disables region proposal networks to reduce detection accuracy. Wei et al. [8] utilized GANs to quickly generate examples. Patch-based attacks apply adversarial patches to image regions to induce incorrect predictions. Compared to pixel perturbations, patches are more physically realizable, as they can be printed or applied in real scenes. Brown et al. [9] introduced adversarial patches targeting classifiers. Liu et al. [10] proposed DPatch for attacking both localization and classification. Lee and Kolter [11] later improved patch update methods for better suppression of detection outputs.

However, existing patch-based methods mainly focus on natural images, and less attention has been paid to physical attacks on ships detection in remote sensing imagery. Moreover, to improve the transferability of black-box adversarial attacks, data augmentation strategy has been widely used [12]–[14]. But current methods apply the strategy to the whole images. This may disturb background regions excessively and result in false detections of the object detection model. Nevertheless, the errors of such false detections are not caused by the adversarial patches themselves but rather by the overaugmentation of background and non-target areas, destabilizing the adversarial patch optimization and weakening attack effectiveness.







Fig. 1: Examples of images after localized data augmentation processing

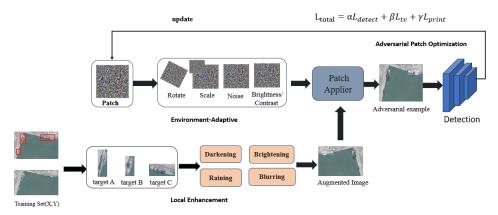


Fig. 2: Framework for generating adversarial patches based on local data augmentation

Focusing on the attacks to ship detection from remote sensing images and revealing the vulnerability of the detection models, this paper propose a adversarial patch generation method via localized augmentation. The proposed method applies transformations only to target regions, as shown in Fig.1, avoiding any influence on non-target areas. By reducing background interference, this approach enables the loss function to focus more directly on the impact of the adversarial patch on the detection model, thereby improving patch optimization stability and attack performance. Using HRSC2016 dataset, experiments on three YOLOv5 variants (YOLOv5-M, YOLOv5-S, YOLOv5-N) demonstrate that the proposed method significantly improves attack success rates. Transferability experiments show that patches generated via localized augmentation exhibit stronger cross-model attack capability, improving robustness in diverse remote sensing scenarios.

II. PROBLEM DEFINITION

Given a clean image I, the attacker aims to optimize a malicious perturbation E, which is added to the original image to produce an adversarial example I', thereby misleading the object detector to make incorrect predictions. In this paper, the generation of the adversarial perturbation E (i.e., the adversarial patch) can be conceptually described as:

$$I' = I \odot (1 - M) + E \odot M$$

$$f(I') \neq f(I)$$
 (1)

where \odot denotes element-wise multiplication, M is the binary mask used to apply the patch, and f represents the victim model whose parameters remain unchanged during the attack. The generated adversarial example I' causes the model to make incorrect predictions, i.e., $f(I') \neq f(I)$.

III. METHODOLOGY

The proposed local augmentation-based adversarial patch generation framework, as shown in Fig. 2, consists of three main modules: local data augmentation, environmental adaptation, and adversarial patch optimization. First, training images are processed by the local data augmentation module,

where target regions are extracted and randomly transformed using operations such as darkening, brightening, simulated rainfall, and blurring, generating target regions under diverse environmental conditions. These transformed regions are then fused with the original images using a mask matrix to form augmented adversarial samples. Next, the environmental adaptation module adjusts factors such as contrast, brightness, and noise to simulate various physical conditions, thereby improving the robustness of the adversarial patches. Finally, the adversarial patch optimization module refines the initial patch by minimizing non-printability score, total variation loss, and detection loss, producing the final adversarial patch.

A. Local Data Augmentation Module

This paper proposes a local augmentation-based data processing method to improve the effectiveness and robustness of adversarial patches. Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i is the input image and y_i is the corresponding target label (including bounding box information), the target regions are first extracted based on the annotations. For each image x, such extraction process can be described as Eq. (2).

$$R = \bigcup_{k=1}^{n} r(y_k) \tag{2}$$

where y_k denotes the annotation of the k-th object, $r(y_k)$ is the pixel coordinate set of the target region computed from the label, R is the union of all target region pixel coordinates, and n is the total number of target objects in the image.

Then, one data augmentation operation such as weather change, blurring, or brightness adjustment is randomly selected and applied to each target region r_j , as shown in Eq. (3), where T_j denotes the randomly selected augmentation operation. Each object is transformed differently to increase diversity.

$$R_{\text{aug}} = \bigcup_{j=1}^{n} T_j(r_j), \quad r_j \in R$$
 (3)

Finally, the augmented regions are placed back to the original image to form the augmented adversarial samples. The process is described as follows:

$$x' = (1 - M) \odot x + M \odot A(R) \tag{4}$$

where M is the mask of the transformed region (1 for target pixels, 0 for background), \odot represents element-wise multiplication, and A(R) is the result of augmenting the original target regions R. This local augmentation strategy avoids background interference introduced by global transformations and better focuses the adversarial optimization on the object area, thus improving attack performance and transferability.

B. Environmental Adaptation Module

Due to varying image quality in maritime remote sensing scenarios, adversarial patches need to adapt to different environments. We introduce a scene intensity matching method by adjusting the contrast, brightness, and noise of the patch to fit the environment characteristics. Based on the printable adversarial patch P, we perform scene adaptation as:

$$P = ((P \times \text{scene}_c) + \text{scene}_b) + \text{scene}_n \tag{5}$$

where scene_c is the scene contrast factor, scene_b is the brightness shift, and scene_n is the added uniform noise. Adjusting these parameters improves patch effectiveness in varying scenes, enhancing the success rate of attacks.

After intensity matching, we apply affine transformations to the patches, including scaling to fit the target bounding box and rotation to simulate viewing angles of ship images. These transformations enhance the robustness of the patches in realworld physical attacks.

C. Adversarial Patch Optimization

To optimize the adversarial patches to degrade the performance of ship detection systems in remote sensing imagery, we adopt a combined loss function that includes non-printability score loss $L_{\rm nps}$, total variation loss $L_{\rm tv}$, and detection loss $L_{\rm det}$.

The **non-printability score loss** L_{nps} measures whether the patches can be realistically printed:

$$L_{\text{nps}} = \sum_{p \in P} \min_{\zeta \in \mathcal{Z}_{\text{print}}} \|p - \zeta\|^2$$
 (6)

where p is a pixel in the patch P, and \mathcal{Z}_{print} is the set of printable colors. This loss minimizes the distance between patch pixels and printable colors to ensure fidelity during physical realization.

The **total variation loss** ensures that the generated adversarial patch maintains visual smoothness, thereby avoiding abrupt pixel transitions and making the patch appear more natural. The total variation Loss is computed as follows:

$$L_{\text{TV}} = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \left((P_{i,j} - P_{i+1,j})^2 + (P_{i,j} - P_{i,j+1})^2 \right)$$
 (7)

where N is the total number of pixels in the patch, and $P_{i,j}$ denotes the pixel value at position (i, j).

To effectively reduce the accuracy of the object detector, the **detection loss** is also included to minimize the maximum confidence score of the target object detected by the model. The loss is defined as:

$$L_{\text{tv}} = \max_{b \in B} S(f(I'), b) \tag{8}$$

where S(f(I'), b) denotes the confidence score output by the detection model f for the bounding box b on the adversarial example I', and B is the set of all detected bounding boxes.

The final optimization objective is a weighted sum:

$$L = \alpha L_{\text{det}} + \beta L_{\text{tv}} + \gamma L_{\text{nps}} \tag{9}$$

where α , β , and γ are hyperparameters that balance attack effectiveness, visual smoothness, and printability.

IV. EVALUATION

A. Datasets

This study conducts experiments on the HRSC2016 dataset, a publicly available high-resolution remote sensing dataset specifically designed for ship detection. HRSC2016 contains remote sensing ship images collected from various scenarios such as ports and open sea areas. The dataset includes 1,061 high-resolution images with a total of 2,976 ship targets. The training, validation, and test sets consist of 436, 181, and 444 images, respectively. The annotations are provided in both horizontal bounding box (HBB) and oriented bounding box (OBB) formats, making it suitable for rotated object detection tasks. Since the image resolutions in HRSC2016 range from 300×300 to 1500×900, we perform data preprocessing before training by removing objects that occupy less than 0.05\% of the image area to ensure a reasonable patch region on each object. The images are resized to 640×640 with gray pixel padding to preserve the aspect ratio.

B. Threat Model

We trained three YOLOv5-based detectors—YOLOv5-N, YOLOv5-S, and YOLOv5-M—on the HRSC2016 dataset. The YOLOv5 [15] series is widely adopted in remote sensing scenarios due to its favorable trade-off between accuracy and computational efficiency. The three models represent small, medium-small, and medium-sized variants, respectively, allowing comprehensive evaluation of patch performance across different model complexity. When measured at an IoU threshold of 0.5, the average precisions are 94.8 (YOLOv5-N), 96.6 (YOLOv5-S), and 96.9 (YOLOv5-M). These results indicate that the detectors trained on the HRSC2016 dataset provide a strong foundation for generating robust adversarial patches.

C. Evaluation Metrics

To evaluate attack effectiveness, we adopt commonly used metrics including average precision (AP), recall (R), and attack success rate (ASR), following the standard definitions in [10].

D. Experimental Setup

In this experiment, the input size of the model is set to 640×640, and training is conducted using Automatic Mixed Precision (AMP). The initial size of the adversarial patch is set to 64×64, with pixel values constrained between 0 and 255. The patch is initially initialized in grayscale mode and then converted to RGB format. During training, the patch position is randomized, and it undergoes various transformations including rotation and translation, with an offset range of [-0.1, 0.1]. Additionally, the target-patch size ratio is set to 0.12, and the patch position is updated in each training iteration. The optimizer used is Adam, with an initial learning rate of 0.03, and the training process spans 200 epochs. This experiment focuses on single-class object detection, specifically targeting the "ship" category.

E. Comparison Experiments

We train adversarial patches separately for YOLOv5-M, YOLOv5-S, and YOLOv5-N, and evaluate their impact on recall, average precision, and attack success rate (ASR) under white-box settings. To assess whether performance degradation stems from adversarial characteristics rather than mere occlusion, we compare the effectiveness of our patches with that of random noise patches. Fig. 3 illustrates the attack effects of our patches on HRSC2016. Table. I presents the average precision, recall, and ASR of each YOLOv5 variant under the attack of adversarial and random patches (IoU=0.5). The optimized adversarial patches significantly compromise detection performance. For example, YOLOv5-S achieves an ASR of 49.3% under adversarial attack—far higher than the 10.5% from random patches—demonstrating that the decline is not merely due to occlusion, but to the adversarial nature of the patch. Furthermore, our method reduces YOLOv5-N's precision from 99% to 30.9% and recall to 42.7%, achieving 58% ASR. These results confirm that the main factor driving performance degradation is the targeted adversarial design, not simple visual obstructions.

TABLE I: Comparison experiment results on YOLOv5 variants using the HRSC2016 dataset.

Model	Patch Type	Average Precision (%)	Recall (%)	ASR (%)
	No Patch	99.0	99.0	0.0
YOLOv5-M	Random Patch	92.7	93.3	7.1
	Adversarial Patch	38.5	42.7	54.1
YOLOv5-S	No Patch	99.0	99.0	0.0
	Random Patch	88.3	89.7	10.5
	Adversarial Patch	38.5	43.7	49.3
YOLOv5-N	No Patch	99.0	99.0	0.0
	Random Patch	88.8	89.9	7.6
	Adversarial Patch	30.9	35.7	58.0

F. Ablation study

This paper proposes to applies transformations locally to target regions instead of the global whole image, to reduce background interference and improve attack performance. To evaluate the effectiveness of the proposed strategy, additional comparison experiments are conducted using YOLOv5-M, YOLOv5-S, and YOLOv5-N: the baseline (no augmentation),

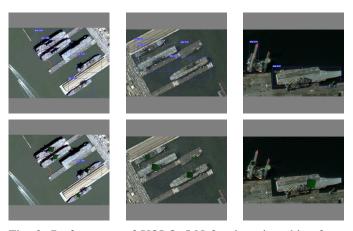


Fig. 3: Performance of YOLOv5-N for detecting ships from the images of HRSC2016 dataset. The first row presents the results from clean images, while the second row shows the results after applying the adversarial patches. It can be clearly seen that after adding the adversarial patches, the ship detector exhibits a significant number of missed detections.

global augmentation, and local augmentation. Table II shows the comparison results.

Compared to the baseline, global augmentation led to mixed results: it improves performance on YOLOv5-M (recall drops by 11.2%, ASR increases by 7.3%), but degrads performance on YOLOv5-S and YOLOv5-N, with ASR on YOLOv5-S even decreasing by 2.5%. In contrast, the local augmentation method consistently outperforms the others. On YOLOv5-M, it reduces recall by 16.6% and improved ASR by 12.1%. On YOLOv5-S, recall drops by 3.5% while ASR increased by 1.1%, again surpassing the global method.

For the lightweight YOLOv5-N, neither augmentation method brought clear benefits. Local augmentation performed similarly to the baseline (recall 35.7% vs. 35.6%, ASR 58.0% vs. 58.8%), and global augmentation slightly improved recall (38.0%) but lowered ASR (56.3%). This suggests that smaller models may lack sufficient feature extraction capacity, limiting the impact of fine-grained local perturbations.

Overall, the results, particularly the improvements on YOLOv5-M, demonstrate that the proposed local augmentation method effectively enhances adversarial patch robustness, especially for medium and large-scale detection models.

TABLE II: Ablation study results on YOLOv5 variants on the HRSC2016 dataset

Model	Method	Recall	Attack Success Rate
YOLOv5-M	Baseline	59.3%	42.0%
	Global Aug	48.1%	49.3%
	Local Aug	42.7%	54.1%
YOLOv5-S	Baseline	47.2%	48.2%
	Global Aug	48.8%	45.7%
	Local Aug	43.7%	49.3%
YOLOv5-N	Baseline	35.6%	58.8%
	Global Aug	38.0%	56.3%
	Local Aug	35.7%	58.0%

G. Transferability Study

To evaluate the transferability of adversarial patches, we conducted a series of experiments in which patches trained on three white-box models were tested on two black-box models. Table III compares the performance of the baseline, global augmentation, and local augmentation strategies across different YOLOv5 architectures.

The results indicate that while global augmentation can enhance transferability in certain cases, it is generally less effective than local augmentation. This is primarily due to the additional background noise introduced during global augmentation, which interferes with the optimization of adversarial patches. In contrast, local augmentation significantly improves the generalization capability of the patches in most scenarios. For example, when YOLOv5-M is used as the white-box model, the attack success rates on YOLOv5-S and YOLOv5-N increase by 4.8% and 5.4%, respectively, compared to only 3.3% and 3.9% improvements achieved by global augmentation. Similarly, when YOLOv5-S serves as the white-box model, local augmentation results in improvements of 2.5% and 2.1%, whereas global augmentation leads to only a 1.1% gain and even a -2.2% drop, respectively. These trends, however, are less evident when YOLOv5-N-a smaller model—is used as the white-box model. This may be attributed to the model's limited feature extraction capacity, which restricts the effectiveness of fine-grained perturbations and diminishes the overall impact of both augmentation strategies. Despite this, local augmentation still performs better than global augmentation, causing only a 1.6% drop on YOLOv5-M and achieving a 1.3% increase on YOLOv5-S. In contrast, global augmentation results in a more substantial performance drop of 5.4% on YOLOv5-M and 1.4% on YOLOv5-S.

TABLE III: Transferability experiment results of attacks across different YOLOv5 architectures.

Source Model	Method	YOLOv5-M	YOLOv5-S	YOLOv5-N
	Baseline	42.0%	33.5%	34.3%
YOLOv5-M	Global Aug	49.3%	36.8%	38.2%
	Local Aug	54.1%	38.3%	49.9%
	Baseline	32.0%	48.2%	47.4%
YOLOv5-S	Global Aug	33.1%	45.7%	45.2%
	Local Aug	34.5%	49.3%	49.5%
	Baseline	32.9%	40.6%	58.8%
YOLOv5-N	Global Aug	27.0%	39.2%	56.3%
	Local Aug	31.3%	41.9%	58.0%

V. CONCLUSION

To generate physically adversarial patches with higher transferability, this paper proposes a local augmentation based adversarial patch generation method, aiming to improve both the attack success rate and transferability of physical attacks in ship detection tasks. By applying local augmentation to only the target region instead of the entire image, interference to the background and irrelevant areas is reduced, thereby improving the stability of the patch optimization process. Experiments based on the HRSC2016 dataset demonstrate that the proposed method can increase the attack success rate of adversarial

patches and enhance their transferability across different object detection models.

Although the proposed method effectively reduces the accuracy of ship detection models, its attack success rate remains relatively low compared to the object detection from natural images. In real-world scenarios, ships are typically large in size, and the adversarial patch is relatively small in comparison, making it difficult to cover the ship's key features. As a result, the interference effect of the adversarial patches is relatively weaker. How to find the better position to apply the adversarial patches will be the focus of our future work. Moreover, this study has not yet conducted physical-world tests of the adversarial patch. Future research will extend to real-world experiments to explore the impact of printed patches on actual systems.

REFERENCES

- Bo Li, Xiaoyang Xie, Xingxing Wei, and Wenting Tang. Ship detection and classification from optical remote sensing images: A survey. *Chinese Journal of Aeronautics*, 34:145–163, 2021.
- [2] Ni Lei, Chunlei Huo, Xin Zhang, Peng Wang, and Zhixin Zhou. Groupnet: Learning to group corner for object detection in remote sensing imagery. *Chinese Journal of Aeronautics*, 35:273–284, 2022.
- [3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [4] Arka Ghosh, Sankha Subhra Mullick, Shounak Datta, Swagatam Das, Asit Kr. Das, and Rammohan Mallipeddi. A black-box adversarial attack strategy with adjustable sparsity and generalizability for deep image classifiers. *Pattern Recognition*, 122:108279, 2022.
- [5] Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Adversarial objectness gradient attacks in real-time object detection systems. In Proceedings of the 2020 IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pages 263–272. IEEE, 2020.
- [6] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1369–1378, 2017.
- [7] Yinpeng Li, Ding Tian, Ming-Ching Chang, Xiaoming Bian, and Siwei Lyu. Robust adversarial perturbation on deep proposal-based models. arXiv preprint arXiv:1809.05962, 2018.
- [8] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. arXiv preprint arXiv:1807.12641, 2018.
- [9] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017.
- [10] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299, 2018.
- [11] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. arXiv preprint arXiv:1904.11897, 2019.
- [12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9185–9193, 2018.
- [13] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 476–485, 2020.
- [14] Jiadong Lin, Chuanbiao Song, Kaichen He, Lizhen Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations* (ICLR), 2020.
- [15] Glenn Jocher, Ayush Chaurasia, Adam Stoken, et al. ultralytics/yolov5: v6.1 - tensorrt, tensorflow edge tpu and openvino export and inference. *Zenodo*, 2022.